DOCUMENT RESUME

ED 042 150 | 24 | AL 002 509

AUTHOR	Hass, Wilbur A.
TITLE	Productivity of Syntactic Forms as a Parameter in Language Development.
INSTITUTION	National Lab. on Early Childhood Education , Chicago, Ill. Early Education Research Center.
SPONS AGENCY	Institute of International Studies (DHEW/OE), Washington, D.C.
REPORT NO	Doc-70706-G-BA-U-28
BUREAU NO	BR-7-0706
PUB DATE	Apr 70
CONTRACT	OEC-3-7-070706-3118
NOTE	12p.; Informal paper
EDRS PRICE	EDRS Price MF-$0.25 HC-$0.70
DESCRIPTORS	*Child Language, *Information Theory, *Language Development, Nominals, *Psycholinguistics, Statistical Analysis, Structural Analysis, *Syntax, Verbs

ABSTRACT

The author raises the question of what one can say about the structure of a person's language from a sample of his speech production and urges the calculating of information theory parameters for grammatical constructions. What has to be done is to decide what construction to focus on and what types to recognize as exemplifying that construction. The author and his colleagues have worked out such breakdowns for three sorts of constructions: finite verb phrases, noun phrases, and classical components, and tallied the occurrences of each of these three sorts of constructions for samples of speech collected from 30 children each (age 5, 6, 7, 9, 11, and 13). For all three sets of construction types there are definite increases with age in the 5-13 year span. It is clear that the older children have to do more selecting when they produce the obvious English syntactic categories of verb phrase, noun phrase, and sentence frame. The author regards the best grammar for a child as that which generates the highest H's (the measure of average surprise value of selections from sets of entities with different probabilities of occurrence) and relative H's for his produced speech. (AMM)

BR-7-0700
24

NATIONAL LABORATORY ON EARLY CHILDHOOD EDUCATION

ED042150

Productivity of Syntactic Forms as a Parameter in Language Development

Wilbur A. Hass
University of Chicago

Informal Paper

AL 002 509

ED042150

Document Number 70706-G-BA-U-28
Printed April, 1970

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

AL 002 509

Productivity of Syntactic Forms as a Parameter in Language Development

Wilbur A. Hass
University of Chicago

Informal Paper

The research or work reported herein was performed pursuant to a contract with the Office of Education, U. S. Department of Health, Education, and Welfare through the Chicago Early Education Research Center, a component of the National Laboratory on Early Childhood Education, contract OEC-3-7-070706-3118.

Contractors undertaking such work under Government sponsorship are encouraged to express freely their professional judgment in the conduct of the work. Points of view or opinions stated do not, therefore, necessarily represent official Office of Education position or policy.

PRODUCTIVITY OF SYNTACTIC FORMS AS A PARAMETER IN LANGUAGE DEVELOPMENT

Wilbur A. Hass, University of Chicago

The title of this symposium, New Problems and Methods in Language Development Research, fits my paper reasonably well. I wish to pose a problem which is general in language development research and to propose a method which indicates a way of working on it. Neither the method nor the problem are actually new, but putting them together, as far as I can tell, is.

First, the problem. As I mentioned it isn't really a new problem, but the problematic nature of it has become inescapably evident only in the course of recent years. It is: what can one say about the structure of a person's language from a sample of his speech production? In particular, how can free speech production be used as a systematic source of data in the study of syntactic development? Recording children's utterances has been a classical activity of parents and psychologists for generations, figuring prominently in the baby biographies, and reaching a peak in the counts of the thirties and forties (so well typified by D. McCarthy's chapter in Carmichael's manual). It continues in the use of indices of complexity, which are still in frequent use, nowadays usually based on some notion of transformedness of sentences (an example of this was given day before yesterday in the Baldwin's paper).

The problem here is that one is never sure that the grammar being applied to the child's language is indeed his grammar. Although this problem has been noticed by perceptive psychologists at least since John Dewey in 1896, it has been fully recognized only since Roger Brown announced the motto that each child should be treated as if he were the speaker of an exotic language which must be analyzed from scratch. The resulting Brownian movement in developmental psycholinguistics (of which a prominent representative, Mrs. Bellugi-Klima, is with us at this table), has applied techniques of structural linguistic analysis to a sizeable sample of the speech of a few growing children, with great dedication and insight. The resulting formulations of processes involved in syntactic development need not be reviewed here. I do need to remark on one limitation; the methods become prohibitively ungainly and perhaps also less enlightening as one approaches the school years. As has often been remarked, children of four-and-a-half do produce speech exemplifying most English syntactic constructions.

How can one use free speech samples to find ways in which syntactic processes may, to cite one obvious intuition (which I will center on today), work differently for five-year-olds than for thirteen-year-olds?

I would like to demonstrate a method--admittedly only one method but the best I have found--for analyzing speech samples. It is based on

the information statistic, H, which (as most of you know without my saying) measures the average surprise value of selections from sets of entities with different probabilities of occurrence, such that predicting whether heads or tails turns up on a coin involves one bit; predicting which of twenty-six letters from an alphabet involves 4,61 bits; and predicting letters according to their frequencies in English involves somewhat less. The general rule is that, in a regular way, increasing the number of entity types increases the value of H and so does divergence of the probabilities of the entity types from equi-probability.

Now, information theory has gotten exceedingly bad press in the psycholinguistic literature lately. People seem to have gotten the idea that information theory indices are inherently tied up with some finite state conception of grammar (some sort of Markoving chaining). Nothing could be further from the truth. Calculating H no more prejudices the kind of grammar one can use than calculating standard deviations on test scores prejudices one toward a particular theory of cognition. In fact, it has long been noted that H is the exact analogue of the standard deviation for nominally scaled data--H tells one how divergent, how much spread out the distribution over classes is--and language data are such nominally scaled data par excellence!

Those of you acquainted with the history of the psychology of language in this country will remember that Wendell Johnson promoted the use of something called the type-token ration (TTR) back in the forties. It is, in fact, a simple version of H. It indicates dispersion, but neglects any feature of the distribution other than number of entity types occurring in a given string of running tokens. It had the additional drawback of being correlated with length of language sample, at least for word types, which was its main use. Although it did have certain use in developmental and stylistic investigations, it has been generally abandoned.

Information theory gained a quick zenith of popularity in the early fifties. It has gained some applicability to sound, letter, and word distributions in constrained and unconstrained language. Thus DiVesta has calculated H for adjectives produced in a constrained association procedure; and Carterette and Jones have computed H and redundancy scores for letters in texts from first, third and fifth grade readers. What I want to urge today is the calculating of information theory parameters for grammatical constructions.

What has to be done is to decide what construction to focus on and what types to recognize as exemplifying that construction. We, at Chicago, have worked out such breakdowns for three sorts of constructions: finite verb phrases, noun phrases, and classical components. In the case

of finite verb phrases, we recognize the following things as leading to different types: past tense inflection; perfective "have"; progressive; passive; negative; imperative; modals; catenative; and emphatic "do." Each finite verb phrase is then scored for which of these markers it contains, each different combination of markers being regarded as a different verb phrase type. In the case of noun phrases, we divide the modifiers into six main classes, three typically occurring before the noun--pre-determiner; determiner; and adjective--and three occurring after the noun--prepositional phrase; participle; adjective clause. Again, each different combination of modifiers is regarded as a different modifier type. In the case of clausal components of sentences, we look at the classical sorts of dependent clauses--noun; adjective; adverb; adverb modifying adjective; etc.--in addition to four types of verbal units--present participle; past participle; gerund; infinitive. The manual in which we have worked out conventions for scoring each of these is available if you want to try your hand at them. We have no illusions about the level of sophistication of the syntactic analyses involved in these breakdowns; we wanted them to be usable by persons with only freshman English and a certain talent and willingness for working on language samples. Our choice of the three constructions was based on the fact that they are of differing "size" and differing number of exemplifying types.

We have tallied the occurrences of each of these three sorts of construction, for samples of speech collected from 30 children each, at the ages of 5, 6, 7, 9, 11, and 13. Under a procedure designed and supervised by Joseph Wepman, each child was encouraged to tell stories to a ten-card TAT; and his productions were recorded and expertly transcribed. (With only slight modification, the techniques should work for any free-speech situation, in which good recordings can be made). We have also analyzed other aspects of the children's lexical and syntactic output, but I'll only be reporting on the information theory statistics today. (We'll be doing the full report this summer.)

We found that the protocols could be scored with good reliability--two independent scorers agree 83 per cent to 97 per cent of the instances, depending on the scorers, the construction being scored, and the age of the child being scored. There seems to be reasonable stability of a given child's indices after at least 100 constructions have been tallied, but our information on this respect comes only from the older children (who might be more stable anyhow).

For all three sets of construction types--verb phrases, noun phrases, and clausal components--there are definite increases with age in the five-thirteen year span. By the general interpretation given H, it is clear that the older children have to do more selecting when they

produce the obvious English syntactic categories of verb phrase, noun phrase, and sentence frame.

We have tried to examine what is involved somewhat more closely, in three ways. First, we have also computed relative H's, that is, the ratio of the obtained H in each case to the maximum H that would be possible given the number of construction types that child used (this is given by log-to-the-base-two of the number of construction types). In this way it is possible to completely rule out the aspect of variability due to the kind of factor measured by the old TTR, and to see how evenly distributed the child's construction instances were over the types he used, however few or many that might have been. When this is done, the resulting age trend for the verb phrases becomes a completely flat one, ranging from .55 to .65 (as a matter of fact, we have just computed the same relative H for the verb phrases of a group of mentally retarded and a group of children with deficient language, and the resulting value is still about .65). Thus the increase with age for verb phrase H is in fact due to the use of more different types of verb phrases for older children. Our evidence does not allow us to rule out the same possibility for the noun phrases (although in this case I still have some feeling that something else is going on). In the case of clausal components, the general increase in H with age remains when one examines relative

H; not only are the older children using more clausal patterns, but they are distributing their use of those that are present more evenly over the different types.

A second technique, which we are using in an attempt to find out more about the noun phrases and verb phrases, is to compute the redundancy which is introduced by looking at where, in the surface structure of the sentence, a given construction is introduced. One would expect that if the construction were the same (in terms of processing) wherever it occurs in the sentence, no redundancy would be created by adding such information. In fact, with both adjectives and prepositional phrases, it makes a great deal of difference whether they occur as a component of a noun phrase or of a verb phrase. This is particularly true of the younger children (although I cannot report a definite age trend, since it has not yet been computed). Our younger children, then, seem to be "context sensitive" to a much greater extent than the older ones.

The last point returns directly to the problem I originally raised. How do we know that the findings we get are not due to forcing an alien grammar onto the utterances of our subjects? As you can tell from the previous point, I think that is indeed evident from some of the findings. What I want to do here in this regard is to make a programmatic statement on the topic. I would regard the best grammar for a child as

that which generates the highest H's and relative H's for his produced speech (at least, I would assert that this is the best evidence of the goodness of a grammar one can get just from having a free-speech sample). I think this proposal is in line with the classical linguistic notion of contrast, which has been applied in Brown's and Erwin-Tripp's analyses of children's language corpora. Finding a contrast between two forms is the all-or-none variant of reporting equal usage of two forms as alternative members of a construction for which H is being calculated. We know from Zipf's counts that we cannot expect even adults' relative H's to be very high; but we also can guess that unless a person can use a form with facility, it will not appear equally as often in his speech as a form which is fully productive for him.

It is in this last sense, that information statistics come to have their most promising use in the analysis of children's speech. They can tell us with respect to what constructions he is acting in a flexible manner, utilizing the resources of the language, and with respect to which ones he is using few possibilities. Information theory measures, I hope, can tell us "where the action is at" in different stages in language development, and with respect to what aspects of language the action is missing in one or another child.

Once given this information, one can go on to design more clever (or, in Chomsky's phrasing, "devious") experimental probes into what the

child is doing. Without the information unconstrained samples of speech can provide when examined in this way, one is likely to miss parts of the growing mastery we call language acquisition.